Interpreting textual distribution: social and situational factors

Stig Johansson

University of Oslo, 2006

As explained in Note 1 on page 1, this paper was originally prepared for a book project which never materialized. Christian Mair refers to it in a paper from 2009 as "forthcoming" (Mair, 2009, "Corpus linguistics meets sociolinguistics: the role of corpus evidence in the study of sociolinguistic variation and change, in A. Renouf & A. Kehoe (eds), *Corpus Linguistics – Refinements and Reassessments*. Amsterdam: Rodopi); it must thus at some point also have been intended for publication in *Arbeiten aus Anglistik und Amerikanistik 34*, but did not appear there.

The present version of the paper was formatted by Hilde Hasselgård and published at http://www.hf.uio.no/ilos/forskning/grupper/Corpus_Linguistics_Group/papers/, but no changes were made to its content apart from updating a couple of obsolete URLs.

Interpreting textual distribution: social and situational factors[1]

Stig Johansson
University of Oslo

Abstract

Problems of interpreting corpus data are considered, with special reference to social and situational features. It is argued that the interpretation of corpus data is intimately bound up with the type of corpus and the nature of the research question. The discussion focuses on publicly available corpora, starting from the Brown and LOB corpora and moving up to more recent corpora, such as the British National Corpus and the International Corpus of English. In conclusion, some suggestions are made for future research.

1. Beyond free variation

There was a time when social and situational factors of language use tended to be overlooked in mainstream linguistics:

> It is customary (except in works devoted specifically to this question) to abstract from synchronic variation in language, either by restricting the description of a language to the speech of a particular group using a particular 'style', or by describing the language in terms of such generality that the description is valid (in intention at least) for all 'varieties'. Some degree of 'idealization' is involved in either of these two procedures, and this may be necessary at the present stage of linguistic theory. (Lyons 1968: 50)

Variation was not denied, but the main focus was on describing language structure in isolation from the conditions of use. With the development of variationist sociolinguistics, pioneered by Labov (1966, 1972), it became increasingly apparent that language use is conditioned to a great extent by social and situational factors and that these cannot be ignored in language description. What some may have rejected as 'free variation' turned out to be highly patterned.

The availability of computer corpora has greatly advanced our knowledge of language variation. Whereas studies within the quantitative sociolinguistic paradigm were often concerned with individual linguistic features elicited in an experimental situation, computer corpora have provided easy access to a vast amount and a broad range of authentic texts. Given these data sources and the associated analysis tools, it has become possible to analyse language variation in great depth and on a scale which was not possible before. Interpreting linguistic variation may not be straightforward, however. In this paper I focus on research on publicly available corpora of present-day English texts, many of which were specifically

[1] This paper was originally prepared for a book project which never materialised. For comments on an earlier version of the paper, I am grateful to Bengt Altenberg, Lund University, my colleagues Johan Elsness and Hilde Hasselgård, University of Oslo, and Geoffrey Leech, University of Lancaster. The version printed here was written in 2006.

designed for language comparison. In conclusion, I suggest some directions for future research.

2. The Brown family

To start with two of the earliest and most influential English corpora, *The Brown Corpus* and *The Lancaster-Oslo/Bergen Corpus* (LOB), these were designed to be representative of printed American and British English texts, respectively. The aim of the LOB Corpus project was to assemble a British English counterpart of the Brown Corpus. Rather than concentrating on limited categories of texts to be used for specific purposes, both corpora were intended to provide a general representation of written text categories for use in research on a broad range of aspects of the language. To facilitate a comparison, an attempt was made to match the British English material as closely as possible with the American corpus. Both contain 500 text extracts of about 2,000 words each, or about a million words in all. The year of publication (1961) and the sampling principles were identical, though there were necessarily some differences in text selection (Johansson *et al*. 1978). The basic composition of the corpora is summarised in Table 1. The two corpora have been extensively used both for inter- and intra-corpus comparisons.

Table 1. The basic composition of the Brown Corpus and the LOB Corpus (differences are marked by italics)

Text categories	Number of texts in each text category	
	Brown	LOB
A Press: reportage	44	44
B Press: editorial	27	27
C Press: reviews	17	17
D Religion	17	17
E Skills, trades, and hobbies	*36*	*38*
F Popular lore	*48*	*44*
G Belles lettres, biography, essays	*75*	*77*
H Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30	30
J Learned and scientific writings	80	80
K General fiction	29	29
L Mystery and detective fiction	24	24
M Science fiction	6	6
N Adventure and western fiction	29	29
P Romance and love story	29	29
R Humour	9	9
Total	500	500

2.1 Comparing LOB and Brown

The Brown Corpus compilers early published a quantitative description of the corpus, focusing in particular on word frequency and sentence length (Kučera and Francis 1967). A similar quantitative account was given after the completion of the LOB Corpus (Hofland and Johansson 1982). Besides presenting information on the LOB Corpus itself, we included lists comparing frequencies in the LOB Corpus and the Brown Corpus as well as some discussion

of differences on various levels: spelling, word form, auxiliary verbs, and some semantic groups of words. Examples:

	LOB	Brown		LOB	Brown
behavior	9	96	*railroad(s)*	1	74
behaviour	119	3	*railway(s)*	66	13
spelled	2	6	*holiday(s)*	107	29
spelt	9	0	*vacation(s)*	3	59
downward	11	16	*baseball*	1	57
downwards	16	0	*cricket*	21	3
he/his/him	17,603	19,412	*man/men*	1,789	2,113
she/her/hers	8,163	6,037	*woman/women*	486	468
shall	348	267	*Mr.*	1,508	839
should	1,276	888	*Mrs.*	292	535

As many of the differences observed confirm what we already knew about British vs. American English language and culture, it served to raise the confidence in the use of the two corpora for inter-corpus comparison also in cases where results were new and unexpected, such as the marked gender bias, with higher frequencies for masculine forms in Brown and for feminine forms in LOB – and a reversal for *Mr*. and *Mrs*. Both corpora, however, agreed in the generally much higher frequencies for the masculine than for the feminine forms, though the contrast was especially striking in the American corpus.

Using the frequency lists of Hofland and Johansson (1982) as a starting-point, Leech and Fallon (1992) further analysed cultural differences between the two corpora, revealing consistent differences in a number of domains: sport, travel and transport, administration and politics, social hierarchy, military, law and crime, business, mass media, science and technology, education, arts, religion, etc. As for the differences with respect to gender-related terms, they note that "the USA was in 1961 already ripe for the feminist movement which hit it in the later 1960s" (p. 43). Summing up the analysis they

> propose a picture of US culture in 1961 – masculine to the point of machismo, militaristic, dynamic and actuated by high ideals, driven by technology, activity and enterprise – contrasting with one of British culture as more given to temporizing and talking, to benefitting from wealth rather than creating it, and to family and emotional life, less actuated by matters of substance than by considerations of outward status. (p. 44f.)

The authors are, however, careful to point out weaknesses of this type of comparison, a topic that will be taken up in Section **2.3**.

The Brown and the LOB corpora have also been used for a large number of studies comparing aspects of grammar: modal auxiliaries, the subjunctive, verb complementation, the genitive, etc.[2] Some studies which include other members of the Brown family will be

[2] See Altenberg (1991) and the ICAME bibliography at: http://icame.uib.no/.

referred to in Section **2.4**. After the Brown and LOB corpora had been tagged for part of speech, it was possible to compare word-class distributions; see Table 2,[3] which shows that there is good agreement between the two corpora. The rank order of the word classes is almost identical, and even the absolute numbers are very close. This contrasts with the considerable differences between the two text category groups of the LOB Corpus. The same picture emerges from a comparison of sentence length (Johansson and Hofland 1989, vol. 1: 17), which reveals that the overall tendencies are very similar, with a consistently higher average sentence length in informative prose. In other words, differences between the two corpora are less marked than those within each corpus. This brings us to intra-corpus comparison.

Table 2. The word-class distribution in the tagged LOB Corpus (A-J: informative prose; K-R: imaginative prose) and the Brown Corpus (total)

Word class	LOB			Brown
	A-J (%)	K-R (%)	Total	
Nouns	26.9	20.0	254,992	272,984
Verbs	16.4	21.9	179,975	185,393
Determiners	13.0	10.5	125,018	123,321
Prepositions	13.1	9.6	123,440	122,613
Adjectives	7.8	5.7	73,546	72,034
Pronouns	5.0	13.1	71,498	66,879
Adverbs	5.0	7.2	56,083	53,283
Conjunctions	5.5	5.4	55,516	60,328
Numerals	2.2	0.9	19,126	20,853
Infinitival *to*	1.5	1.7	15,837	15,030
WH-words	1.5	1.6	15,718	14,921
Not	0.6	1.1	7,454	6,976
Existential *there*	0.3	0.3	2,794	2,280
Interjections	0	0.4	1,109	629
Other	1.3	0.7	11,631	–

2.2 Intra-corpus comparison

Just a glance at Table 1 above makes it clear that there is a need for caution in using the two corpora for intra-corpus comparison, most obviously because the text categories vary greatly in size and some of them are very small indeed.[4] Nevertheless, an intra-corpus comparison may be revealing, at least in connection with common features of the language. Table 2 above shows that there are considerable differences in word-class distribution between the two subsets of the LOB Corpus. In informative prose the relatively more common types were nouns, determiners, prepositions, adjectives, and numerals. In imaginative prose they were verbs, pronouns, adverbs, *not*, and interjections. Most of the differences are evidence of greater complexity at the noun phrase level in informative prose. In imaginative prose the verbs are more prominent (and elements accompanying verbs: adverbs, *not*); the verbs, in

[3] The table, with some simplifications, is quoted from Johansson and Hofland (1989, vol. 1: 15). The figures for the Brown Corpus are quoted from Francis and Kučera (1982).

[4] Oostdijk (1988: 17) observes that "corpora which have being [*sic*] compiled with the intention of representing a cross-section of a language are not suited for the study of linguistic variation since, in selecting a great many samples, they neutralize any variety-specificity."

fact, outnumber the nouns. There are also considerable differences in word class sequences between the two category groups (Johansson and Hofland 1989, vol. 2: 3); thus, for example, the sequence 'singular noun plus singular noun' is more than twice as frequent in informative prose as in the fiction categories.

An intra-corpus comparison of the most frequent words may also be revealing (Hofland and Johansson 1982: 340ff.). For example, the definite article – the most frequent word in the English language – has a higher relative frequency in all the informative prose categories than in the categories of imaginative prose. To take an example of the opposite relationship, the modal auxiliary *could* is relatively more frequent in all the categories of imaginative prose than in the categories of informative prose. For less frequent words as well, an intra-corpus comparison may reveal interesting patterns. Tables 3 and 4 give some plus-words for text categories J (learned and scientific writings) vs. K-R (fiction), more specifically the 40 nouns (except proper nouns and abbreviations), lexical verbs, adjectives, and adverbs with the highest 'distinctiveness coefficient' in each group.[5] A number of observations can be made here, such as: the nouns in J are predominantly abstract, those in K-R are concrete; the verbs in J include a variety of verb forms and process types, those in K-R are almost all *ed*-forms and denote material or behavioural processes. The differences between the adjectives and the adverbs are equally striking.

Table 3. Plus-words in categories J vs. K-R of the LOB Corpus: nouns and lexical verbs. The words are listed in order of their distinctiveness coefficient.

Nouns		Lexical verbs	
J	K-R	J	K-R
constants	mister	measured	kissed
axis	sofa	assuming	heaved
equations	wallet	calculated	leaned
oxides	cheek	occurs	glanced
equation	living-room	assigned	smiled
theorem	cafe	emphasized	hesitated
coefficient	wrist	obtained	exclaimed
ions	darling	executed	murmured
correlation	sigh	tested	gasped
electrons	gun	corresponding	hurried
impurities	gaze	vary	flushed
oxidation	clip	bending	cried
parameters	fist	varying	eyed
nickel	trail	loading	staring
electron	lounge	measuring	paused
impurity	cheeks	determine	whispered
diagram	lips	isolated	waved
ion	cigarette	dissolved	nodded
parameter	stairs	resulting	frowned
coefficients	footsteps	defined	shivered
oxygen	dad	occur	muttered
sodium	lawn	stressed	stared
equilibrium	receiver	illustrates	flung
oxide	madam	recognized	grinned
variable	jacket	identified	laughed

[5] The tables are quoted from Hofland and Johansson (1982: 28f.). As for the calculation of the distinctiveness coefficient, see p. 14, *op.cit*.

evaporation	fool	testing	shrugged
contamination	pistol	follows	jerked
approximation	envelope	observed	tapping
alloy	shoulders	tend	laughing
hydrogen	door	demonstrated	swung
ratios	forehead	exposed	pretended
data	phone	containing	leaning
component	knees	deposited	wondered
symmetry	tears	using	shook
curve	bedroom	forming	kiss
displacement¨	fingers	indicates	straightened
computer	patch	examine	rang
cells	skirt	associated	sounded
curves	eyes	indicate	gripped
particle	pocket	obtain	smiling

Table 4. Plus-words in categories J vs. K-R of the LOB Corpus: adjectives and adverbs. The words are listed in order of their distinctiveness coefficient.

Adjectives		Adverbs	
J	K-R	J	K-R
thermal	damned	theoretically	impatiently
linear	asleep	significantly	softly
radioactive	sorry	approximately	hastily
structural	gay	hence	nervously
finite	miserable	relatively	upstairs
transient	dear	respectively	faintly
physiological	silly	commonly	quietly
numerical	empty	separately	abruptly
magnetic	stiff	consequently	eagerly
conceptual	dreadful	similarly	upright
residual	afraid	rapidly	tomorrow
differential	deadly	thus	downstairs
stationary	sweet	furthermore	gently
statistical	ashamed	sufficiently	anyway
negative	lovely	therefore	maybe
relative	faint	secondly	swiftly
experimental	calm	ultimately	presently
theoretical	silent	readily	suddenly
integral	nice	effectively	somewhere
mechanical	funny	generally	back
chemical	worried	widely	slowly
internal	tired	strictly	desperately
initial	stupid	mainly	sharply
reliable	polite	directly	away
significant	savage	partly	barely
continuous	quiet	previously	backwards
relevant	tall	specifically	somehow
prior	lonely	chiefly	utterly
intermediate	glad	presumably	aboard
liquid	damp	closely	down
equal	dark	accordingly	lightly

rapid	mad	frequently	quickly
constant	pretty	however	inside
imperial	quick	moreover	carefully
consistent	pink	nevertheless	again
positive	clean	unfortunately	off
upper	sudden	briefly	then
aesthetic	desperate	considerably	never
statutory	loud	purely	sooner
external	ugly	originally	scarcely

An early ambitious undertaking designed to study intra-corpus variation is Alvar Ellegård's Syntax Data Project (Ellegård 1978). Samples from the Brown Corpus, more exactly 16 texts from each of text categories A, G, J, and N, were analysed manually on three levels – clause structure of sentences, constituent structure of clauses, and word class of individual words – with the aim of providing "an as nearly complete as possible parsing" (p. 1). Among the findings we note:

> The language seems to be remarkably stable as regards such features as the number of clauses per sentence, and the depth of embedding for such clauses. We might have expected the popular texts to have much shorter sentences (in terms of clauses) and much less embedding than the literary and scientific ones. There is indeed a slight tendency in this direction, but it is very weak. (p. 76).

The distribution of clause types is also quite similar, though one might have expected fewer subordinate clauses in the popular texts. The stability is confirmed by the striking similarity between half-samples from the same text category. The most notable differences have to do with the distribution of word classes – e.g. more verbs and fewer nouns in the popular texts – and the length and complexity of phrases – lowest in the popular texts (N), highest in the scientific texts (J), with the journalistic (A) and literary texts (G), as in most cases, occupying the middle ground.

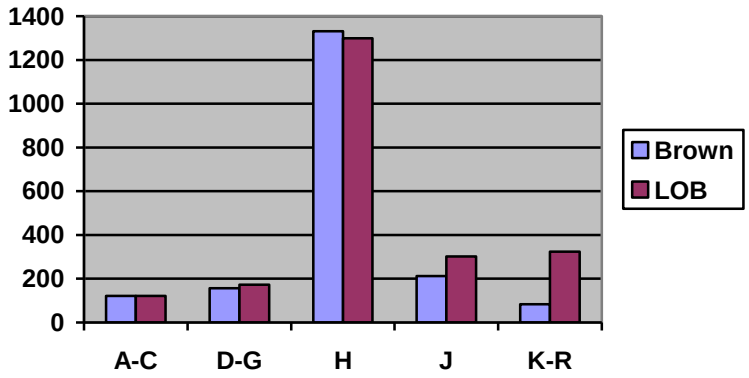

Figure 1. The distribution of *shall* in text categories of the Brown and LOB corpora (relative frequency per million words)

As a final example of intra-corpus comparison, consider the distribution of *shall* in the LOB and Brown corpora; see Figure 1.[6] In spite of the overall difference between the corpora (see **2.1**), the distribution across text categories is very similar: *shall* in category H accounts for the majority of the examples and is about equally common in both corpora; the main difference is found in fiction, where the frequency is much higher in the LOB Corpus, though it is far lower than the figure for category H.[7] In other words, again we see that differences between the two corpora are less marked than those within each corpus.

2.3 Provisional conclusion

If we want to use the two corpora for conclusions on the relationship between British and American English, there is a need for caution. In the first place, the corpora are quite small and contain relatively short text samples, which limits their usefulness for studies of lexis and discourse patterns. Second, they lack many important text categories, notably different types of speech. Third, as we have seen, there may be greater differences within each corpus than between the corpora. A study of textual distributions between the corpora should therefore take intra-corpus variation into account. An example of such a study is Junsaku Nakamura's quantitative comparison of modals in the LOB and Brown corpora (Nakamura 1993).[8] Interpreting textual distributions without taking intra-corpus variation into account is hazardous.

Differences between corpora are often quantitative rather than absolute. For this reason, there is a need for statistical testing. Many of the items in our lists comparing words in the LOB and Brown corpora (Hofland and Johansson 1982: 471ff.) include chi-square values. We stress, however, that these should only be regarded as a rough guide:

> The only thing we can be certain of is that the absence of a significant chi-square value is a good indication that a difference is accidental. It is far more doubtful whether a significant chi-square value justifies safe conclusions about frequency differences. The reader is encouraged to look for consistent behaviour of related words. (p. 39)

Kilgarriff (2001a) has shown the problems in using the chi-square test for this kind of comparison. There is no doubt a need for better statistical measures, but statistical significance is not enough in itself. Are differences linguistically important? Can they be given a reasonable linguistic interpretation? Here it is useful to look for consistent behaviour across individual observations, as suggested in the quotation above. A simple example quoted from Hofland and Johansson (1982: 40) is:

	LOB	Brown		LOB	Brown
firstly	14	0	*thirdly*	10	1
secondly	29	5	*fourthly*	3	0

Although the figures are low, there is a consistent difference between the two corpora. Leech and Fallon's cultural comparison (**2.1**) provides other examples, as do the lists in Tables 3 and

[6] The frequencies for *shall* given here and in the comparisons below include the form *shan't*. There were very few occurrences of *shan't*. Most of them were found in LOB. The few instances of *shalt* (found in quotations) are not included.
[7] For a more detailed account, see Krogvig and Johansson (1984).
[8] Note also the recent paper by Wilson (2005), where there is a similar quantitative analysis of modals.

4 above, in the latter case to do with intra-corpus comparison. By itself, a single word may not mean much; when viewed together, they build up an interpretable pattern.

Another aspect of the comparison of word frequencies, as found in Hofland and Johansson (1982), is that it focuses on form rather than meaning. Yet we know that homonymy and polysemy are prevalent in the language. Interpreting the distribution of forms, without considering meaning, requires great caution. Ideally, we should study forms in context, a point which applies both to lexis and grammar. It is a problem that the use of electronic corpora may tempt one to look only at frequency distributions, ignoring context, or at concordances, without consulting the wider context. This is not a flaw of the corpus itself, but rather of the way it is used. Electronic corpora make it possible for the researcher to examine both macro- and micro-level patterns, varying the analysis according to the focus of the study.

2.4 Extending the family

The Brown and LOB corpora have served as models for other corpora; see Table 5. A look at the dates reveals that there is a difference in the time of publication of the texts. In all cases, however, the compilers of the new corpora attempted to follow the Brown and LOB sampling principles as closely as possible. Nevertheless, there were necessarily some differences. Shastri, the compiler of *The Kolhapur Corpus*, points out that the fiction categories differ from those of the LOB and Brown corpora "because of the inherent difference in the Indian situation" (Shastri 1988: 17). The categories 'Science fiction', 'Adventure and Western fiction', and 'Romance and love story' are much smaller, and more texts are included in the 'General fiction' category (K). According to Sigley (1997: 211), the sample of imaginative writing in WCWNZE is "a single category of general fiction, not subdivided into specific genres (owing to the paucity of special-genre fiction written, edited and published within New Zealand)". Sand and Siemund (1992) report that there were difficulties in matching the texts of the press categories of FLOB with those of LOB. Interestingly, they point out that they "ranked the comparability of LOB '91 to LOB '61 higher in priority than the possible alternative goal, viz. to create the accurate picture of the British printed press right now" (p. 120).

Table 5. The extended Brown family

Corpus	Abbrev.	Date of corpus texts
Australian Corpus of English	ACE	1986
Brown Corpus	Brown	1961
Freiburg Brown	Frown	1992
Freiburg LOB	FLOB	1991
Kolhapur Corpus	-	1978
Lancaster-Oslo/Bergen Corpus	LOB	1961
Wellington Corpus of Written New Zealand English	WCWNZE	1986

In the building of comparable corpora, there is necessarily a tension between comparability and representativeness.[9] In the case of the LOB and Brown corpora, it was

[9] See the discussion in the recent paper by Leech and Smith (2005: 87ff.), where the issue of representativeness vs. comparability is raised in connection with the planning of Lancaster 1931, a new corpus designed to match

generally possible to sample in a corresponding manner, though there were problems, such as finding British texts to match American western fiction. In the case of the other members of the Brown family, problems were greater, not to speak of the considerably more serious problems which must have arisen in matching the corpora of the ICE family (see **6.1**). The tension between comparability and representativeness is something we have to live with, but it means that the researcher must take great care in interpreting the results of corpus comparison. To what extent are differences due to a sampling bias? To what extent can corpus findings be regarded as representative of the varieties the corpora are intended to represent?

As an example, consider first the distribution of *shall* in the Kolhapur Corpus, as compared with the Brown and LOB Corpora. The overall distribution is:

Brown	269	LOB	354	Kolhapur	364

The frequency for *shall* is higher in the Kolhapur Corpus than in the other corpora, as pointed out by Shastri (1988: 18), and he continues:

> This may be due to the predominance of written language over spoken in the Indian pedagogical context. Also, English in India, taught as a second language, tends to retain some of the older usages which might have lost currency in the first language situation.

If we compare the distribution across text categories, we find that the overall tendency is the same as in LOB and Brown; but *shall* is considerably more common in Category H; see Figure 2. This category needs to be examined in detail, before any definite conclusions can be drawn.

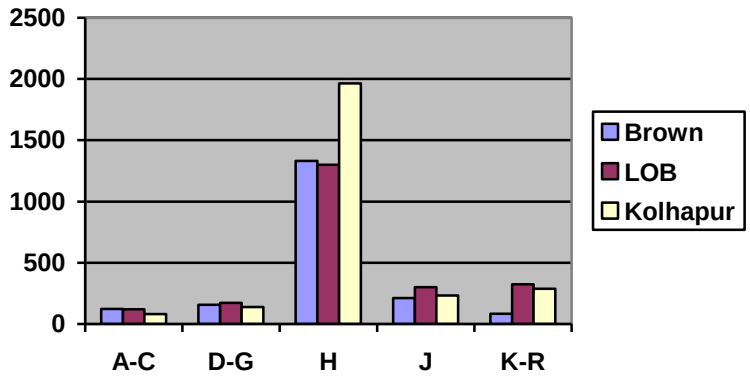

Figure 2. The distribution of *shall* in text categories of the Brown, LOB, and Kolhapur corpora (relative frequency per million words)

If we move on to a comparison of the Brown, Frown, LOB, and FLOB corpora, we find the following distribution of *shall*:

Brown	269	LOB	354
Frown	149	FLOB	197

LOB and FLOB. It is pointed out that "[t]he next planned extension to our project will be a further prequel of this kind, that is, a matching corpus of British English texts published in 1901, provisionally named Lancaster 1901" (p. 84). As the time span is increased, the sampling difficulties will multiply.

For both the British and the American corpora, there has been a reduction in the use of *shall* in the period from 1961 to 1991-1992. Moreover, the gap between the British and American corpora has grown smaller. The latter might be due to a trend towards Americanisation in British English. But how could we explain the decline in both American and British English?

At this point, we turn to the recent paper by Geoffrey Leech (2004), where he examines the distribution of the modal auxiliaries, including *shall*, in the Brown, Frown, LOB, and FLOB corpora. This paper is a model with respect to the care which is taken in analysing and interpreting the data. Leech lists some 'hazardous assumptions: from data description to language description' (p. 70). These deserve to be quoted in full:

1. That the corpora are large enough and varied/balanced enough to allow us to extrapolate from corpus findings to what is happening in (relevant varieties of) the language in general.

2. That the corpora are sufficiently comparable in terms of samples of the varieties represented, and in using the same sampling methods.

3. That statistically significant results can be attributed to real linguistic differences, rather than to extraneous factors such as cultural shifts or faulty sampling.

4. That the grammatical categories are defined and used in a way that other grammarians or linguists find reasonable.

5. That the extraction of data from the corpora has been acceptably (if not totally) free from error.

Leech finds that there has been a general reduction in the frequency of the modals both in the British and the American corpora, though individual modals have been declining at different rates. A possible explanation of changes in FLOB and Frown as compared with LOB and Brown might be a trend towards colloquialisation of written English, as suggested in papers by Christian Mair (1997, 1998). The study is extended to changes which might be indicative of a development towards a more colloquial style, and there are indeed some striking findings: an increase in the use of the present progressive, of verb contractions, verbless questions, the genitive, etc. But how could we account for the process of colloquialisation? For this we need to turn to theories of the cognitive and social workings of language.

Before leaving Leech's important paper, we need to return to *shall* in the Brown, Frown, LOB, and FLOB corpora. Leech suggests that "it is in the nature of corpus research to be provisional" (p. 75), and he mentions on the same page that "[o]nce the gross frequency changes have been plotted, the next step is to investigate factors internal to the corpora that might help explain these changes (e.g. differential results in the different subsections of the corpus)". A study of the distribution of *shall* shows that the greatest changes have taken place in category H, which has been reduced by more than half between the two generations of corpora, although it still stands out as the most frequent of all the text categories.

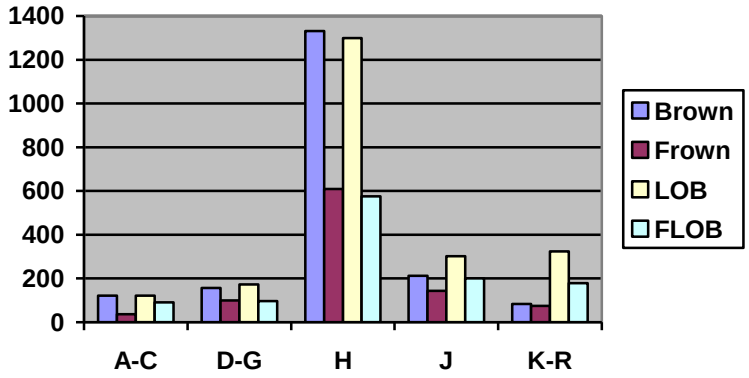

Figure 3. The distribution of *shall* in text categories of the Brown, Frown, LOB, and FLOB corpora (relative frequency per million words)

A closer study of *shall* within category H reveals that the form is very unevenly distributed, which is not unexpected in this heterogeneous category; see Figure 4.[10] The majority of occurrences are concentrated in a limited number of texts. In the LOB Corpus three texts account for 78 of the 95 examples;[11] in FLOB a single text accounts for 32 of the 43 examples;[12] in the rest of the texts there are just scattered instances (no more than three in each). What this shows is that it is necessary to examine single texts closely before firm conclusions can be drawn on developments. A general trend towards colloquialisation is not sufficient to account for the data (see further Section **5**). What sorts of texts are affected, and in what ways are they changed?

Figure 4. WordSmith dispersion plot for *shall* in Category H of Brown (first line), LOB (second line), Frown (third line), and FLOB (last line).

There have been a number of studies comparing members of the extended Brown family, including Hundt (1997), Mair (1997, 1998), Holmes and Sigley (2002), and Mair and Hundt (2002); note the survey articles on Australian English by Collins and Peters (2004) and on New Zealand English by Hundt *et al*. (2004). Many notable patterns are revealed, e.g. this one for the decline of *shall* (quoted from Hundt *et al*. 2004: 589):

```
                BrE            AmE            AusE           NZE
----------------|--------------|--------------|--------------|-------------->
conservative                                                  innovative
```

[10] Each vertical bar indicates a hit, going from the first to the last text. Where hits cluster closely together, they cannot be singled out individually.
[11] These are official government documents: H13, H14, and H24; see Johansson *et al*. (1978).
[12] This is again an official government document: H14, *Agreement Between the Government of the United Kingdom of Great Britain and Northern Ireland and the Government of the Italian Republic Concerning Mutual Assistance in Relation to Traffic in Narcotic Drugs or Psychotropic Substances and the Restraint and Confiscation of the Proceeds of Crime.* Rome, 16 May 1990. London: HMSO. 1991. Pp. 3-7.

Australian and New Zealand English are thus leading the field in the decline of *shall*. How and why the changes have come about, however, remains something of a puzzle which cannot be solved without a close examination of the distribution in text categories and a close study of individual texts.

3. Speech vs. writing

A comparison based on the Brown family can only deal with written or, rather, printed texts. What happens if the comparison is extended to speech? A simple comparison of frequency lists shows that there are great differences. Hofland and Johansson (1982) compare the most frequent words in the LOB Corpus and three other corpora: the Brown Corpus, an American corpus of written texts 'to which students are exposed in school grades 3 through 9 in the United States' (Carroll *et al*. 1971: xiii), and a British corpus of spontaneous conversation (Jones and Sinclair 1974); see Table 6. The spoken corpus stands out as being strikingly different, not just in relation to the two American corpora, but also as compared with the other British corpus. In fact, the closest correspondences are found between LOB and Brown, not between the British or the American corpora.

Table 6. The 50 most frequent words in the LOB Corpus as compared with the ranks of the corresponding words in three other corpora (quoted from Hofland and Johansson 1982: 19; - indicates that a word is not found among the 50 most frequent words in the corpus)

Form	LOB	Brown	Carroll *et al*. (1971)	Jones and Sinclair (1974)
the	1	1	1	1
of	2	2	2	8
and	3	3	3	3
to	4	4	5	6
a	5	5	4	5
in	6	6	6	9
that	7	7	9	11
is	8	8	7	12
was	9	9	13	14
it	10	12	10	7
for	11	11	12	26
he	12	10	11	21
as	13	14	16	40
with	14	13	17	47
be	15	17	21	33
on	16	16	14	23
I	17	20	24	2
his	18	15	18	-
at	19	18	20	35
by	20	19	27	-
had	21	22	29	-
this	22	21	22	20
not	23	23	30	-
but	24	25	31	17
from	25	26	23	-
have	26	28	25	25

are	27	24	15	36
which	28	31	41	-
her	29	35	-	-
she	30	37	-	-
or	31	27	26	27
you	32	33	8	4
they	33	30	19	15
an	34	29	39	-
were	35	34	34	-
there	36	38	37	37
been	37	43	-	-
one	38	32	28	28
all	39	36	23	29
we	40	41	36	50
their	41	40	42	-
has	42	44	-	-
would	43	39	-	-
when	44	45	35	-
if	45	50	44	39
so	46	-	-	38
no	47	49	-	18
will	48	47	46	-
him	49	42	-	-
who	50	46	-	-

The availability of spoken corpora has greatly advanced our knowledge of spoken English and of features which differentiate speech and writing. A major breakthrough was the completion of *The London-Lund Corpus* (Svartvik 1990), which includes a wide range of spoken material collected and transcribed according to the plan for the Survey of English Usage, a project which was designed to collect a corpus of spoken and written texts for use in the description of English (Quirk 1960). Here we are only concerned with the spoken material, which was prepared for computer analysis at Lund University under the direction of Jan Svartvik.[13] Like the members of the Brown family, the London-Lund Corpus (LLC) was made available to the community of researchers through ICAME.[14]

Shortly after the LLC was completed, Bengt Altenberg and Gunnel Tottie initiated a project comparing aspects of speech and writing on the basis of LOB and the LLC. Equal amounts from two sub-varieties were selected, spontaneous conversation from LLC and expository prose from categories D-J of LOB, which could be regarded as the archetypal forms of the two media. In the Introduction to the collection of papers from the symposium which marked the end of the project, we read:

> We […] felt that by restricting our samples to spontaneous conversation and expository prose, we could achieve, at the same time, maximum contrast and maximum comparability – maximum contrast because of the differences in medium as well as formality, and maximum comparability because the speakers participating in the recorded conversations were all educated to academic level, and could be thought

[13] The original London-Lund Corpus consisted of 87 texts of 5,000 words, 34 of which – representing surreptitiously recorded conversation – were published in printed form in Svartvik and Quirk (1980). The 13 texts which were missing at the outset were added later, after being processed at the Survey of English Usage 'in conformity with the system used in the original London-Lund Corpus' (Greenbaum and Svartvik 1990: 14).
[14] See: http://icame.uib.no/.

> of as potential readers or writers of the type of written material included. (Tottie and Bäcklund 1986, Introduction by Tottie, p. 8)

Judging by the findings reported in the papers, this was a good starting-point. Altenberg (pp. 13-40) writes about contrasting linking and focuses on two links where there are marked differences between the corpora: *but* and *anyway*. Drawing on illustrations from the LLC, he shows how *but* can be used for a variety of communicative purposes, such as interactive countering, topic shifting, and topic resumption. Bäcklund (pp. 41-55) examines conjunction-headed abbreviated clauses (e.g. *beat until stiff*), showing that they are used somewhat differently in the two media, possibly related to the difference in the communicative situation. Hermerén (pp. 57-91) surveys the means for expressing modality, including both modal auxiliaries and other structures, paying attention both to form and meaning. Interestingly, he finds that the spoken material is more modal overall, though meanings are differently distributed in the two corpora. In her study of adverbials of focusing and contingency, Tottie (pp. 93-118) arrives at the conclusion that the former are more typical of writing and the latter of speech, which may reflect differences in communicative constraints and situational and communicative needs. Karin Aijmer (pp. 119-129) sets out the answer the question 'Why is *actually* so popular in spoken English?' and Anna-Brita Stenström (pp. 149-163) asks 'What does *really* really do? Strategies in speech and writing'.

Although frequency figures like those given in Table 6 above are suggestive, they need to be further analysed and interpreted. In the papers I have referred to, quantitative observations, such as the finding that *actually* is ten times more frequent in the LLC than in LOB, are merely a starting-point. What strikes the reader is the focus on meaning and function, and the attention to language use in context. Repeated reference is made to Chafe's (1982) description of informal speech vs. formal written language as characterised by fragmentation vs. integration and involvement vs. detachment. This gives added significance to the individual observations and provides a general framework within which the findings can be interpreted.

The availability of the LLC led to a spate of studies analysing spoken English or comparing spoken and written English. Apart from those already mentioned, we find important works such as Altenberg's paper on causal linking in spoken and written English (1984) and the monographs on cleft constructions and negation by Collins (1991) and Tottie (1991), respectively. The analysis of language variation had taken a great step forward.

4. Beyond individual features

Douglas Biber's pioneering work represents another important step forward in the study of linguistic variation. Here I will focus on the research reported in his monograph on variation across speech and writing (Biber 1988). The two main distinguishing characteristics are (1) that he takes into account many categories of spoken and written texts and (2) that the study includes a wide range of linguistic features rather than selected individual forms or structures. This is done in a highly innovative manner which takes maximum advantage of computational and statistical techniques. A similar study would be inconceivable without access to electronic corpora.

Although Biber's work is very well known, I will give a brief outline of his methodology. The study includes 17 written 'genres', viz. texts from the 15 LOB categories and 2 collections of letters, and 6 'genres' from the LLC, all in all about a million words. On the basis of previous work on spoken/written differences, Biber identified 67 features for inclusion in the analysis: tense and aspect markers, place and time adverbials, pronouns and

pro-verbs, questions, nominal forms, passives, stative forms, subordination features, etc. Frequencies were calculated for each feature in each text and normalised to a text length of 1,000 words. A factor analysis of the co-occurrence of features revealed seven factors underlying the variation across the texts. On the basis of the features, both positive and negative, Biber interpreted the factors as representing different textual dimensions (one of the seven factors was not considered strong enough for interpretation): Dimension 1 'Involved versus Informational Production', Dimension 2 'Narrative versus Non-Narrative Concerns', Dimension 3 'Explicit versus Situation-dependent Reference', Dimension 4 'Overt Expression of Persuasion', Dimension 5 'Abstract versus Non-Abstract Information', Dimension 6 'On-Line Informational Elaboration'. Dimension scores were calculated for each text by summing up the frequencies of features associated with each dimension. Mean values were then calculated for each genre such that they could be ranked in relation to each other for each dimension. This reveals, for example, that telephone conversations and official documents are at the extreme ends for Dimension 1, with the other genres placed at different positions along the scale. Relations along each dimension were discussed and illustrated by short text samples.

The most significant finding is that there is no dimension which unequivocally characterises speech vs. writing:

> This analysis shows that there is no single, absolute difference between speech and writing in English; rather there are several dimensions of variation, and particular types of speech and writing are more or less similar with respect to each dimension. (p. 199)

The relationship is multi-dimensional. Thus, for example, official documents are at the bottom for Dimensions 1 and 2, at the top for Dimension 3, in the middle for Dimension 4, at the top for Dimension 5, and fairly low for Dimension 6. In contrast, telephone conversations are at the top for Dimension 1, fairly low for Dimension 2, towards the bottom for Dimension 3, in the middle for Dimension 4, at the bottom for Dimension 5, and fairly low for Dimension 6. Different genres can therefore be similar, or differ, in a variety of respects. Using the same methodology it is possible to show the range of variation within each genre and relationships between sub-genres.

There is no doubt that Biber's work has provided significant new insight and has given us a new methodology which has proved to be seminal and has inspired many researchers to follow suit. Biber himself and his co-workers have applied the methodology to many different types of studies (see e.g. Biber and Finegan 1991, Biber *et al*. 1998): research articles in different fields, relationships among the sections of research articles, student speech and writing, textbooks, language acquisition, historical change, individual author styles, etc.[15] Whereas the focus of earlier corpus-based studies of variation was on individual features, we now have a way of characterising texts as well as text categories and their relationships.

Considering the achievement, it seems petty to draw attention to possible shortcomings.[16] Perhaps the most obvious problem is that the focus is on surface features which can be relatively easily identified and quantified. Even so there may be difficulties in capturing the relevant features. There is a revealing footnote on this in Biber (1986: 388), part of which is quoted here:

[15] The method has even been applied for cross-linguistic comparison (Biber 1995).
[16] For a critical discussion, see Ball (1994).

> […] the goal of the programs was to capture 70-90% of the occurrences of a construction, with no obvious skewing in one mode or another.

Many occurrences of features which are part of the analysis may thus be missed. Moreover, as pointed out by Esser (1993: 54f.), "the possibilities of homonymy and synonymy are excluded, i.e. the possibility that a linguistic form may serve more than one communicative purpose". Equally important, as Biber (*loc.cit.*) himself admits, "Other features were not included because they cannot be analysed automatically – e.g. conjoined phrases and conjoined clauses […] and features representing different types of cohesion and information structure". Prosody is not included. Syntactic complexity can only be observed indirectly. Another relevant point is that the texts on which the analysis is based are not full texts but fairly short samples. All of this may cause problems for the interpretation of the dimensions.[17]

To conclude, the approach is very valuable, but must be applied with caution, taking possible limitations into account. What we can hope for is that Biber-type studies will be combined with analyses of the behaviour of specific features and, not least, with micro-level studies of individual texts. As we learn more about these matters, the multi-feature analysis may become more refined.

5. A corpus-based grammar

Biber's variation studies led him to launch a new major undertaking which resulted in a corpus-based grammar, *The Longman Grammar of Spoken and Written English* (Biber *et al*. 1999). Whereas the focus in his earlier work was on how a multi-feature analysis can throw new light on texts and text categories, the grammar project aimed at showing how a study which takes texts and text categories as a starting-point can yield new insight into the use of forms and grammatical structures. Simplifying, we can perhaps say that the question in the former case was: what does the distribution of grammatical features reveal about dimensions of variation?; and in the latter: what does the distribution in 'registers' reveal about grammar? But the most obvious difference is that the grammar takes up a much wider range of grammatical features.

Four main text categories, or 'registers', were identified: news reportage, academic prose, fiction, and conversation. Each can be characterised with reference to major situational differences (p. 16): mode (spoken vs. written), interactiveness and online production, shared immediate situation, main communicative purpose/content, audience, and dialect domain. Large corpora were compiled for each of these registers and analysed for a large number of grammatical features, with a view to revealing how grammatical choices are made and to what extent they differ across registers. Though the analysis is quantitative, there is a consistent attempt in each case to interpret the findings in relation to the situational variables and other relevant factors.

To take a simple example, the type-token ratio (TTR) – incidentally, a point which is usually not dealt with in grammars – is quite different in the four main registers. It is consistently lower in conversation than in the written registers, and somewhat lower in academic prose than in fiction and news reportage. The suggested explanation is:

> TTR is low in conversation because it is less concerned with the transmission of information than writing. Moreover, conversation is spontaneously produced, with

[17] Problems may show up in the naming of the dimensions. Thus, Dimension 5 is named 'Abstract versus Non-Abstract Information' in Biber (1988) and 'Impersonal vs. Non-impersonal Style' in Biber *et al*. (1998: 155).

> little time for planning and varying the choice of words. Repetition is characteristic of spoken language. It may be used for emphasis, to help the planning of the speaker, or to make sure that the message gets across to the hearer. […] The TTR differences among the written registers are more surprising. We naturally expect a somewhat higher TTR for fiction, where the focus is more on form and elegance of expression. The high TTR in news reflects the extremely high density of nominal elements in that register, used to refer to a diverse range of people, places, objects, events, etc. At the other extreme, academic prose has the second lowest TTR, reflecting the fact that a great deal of academic writing has a restricted technical vocabulary and is therefore less variable than fiction and news reportage. (p. 53f.)

Similar patterns are noticed, commented on, and illustrated in all the chapters. By corpus analysis, we can discover patterns which were previously unknown as well as document phenomena which we had suspected. The distributions show a remarkable degree of consistency, and it is usually possible to provide plausible interpretations for the patterns observed. The most notable part is perhaps the final chapter on the grammar of conversation which brings together, interprets, and expands on observations which are made throughout the book.

At the same time, there is a need for caution. Many grammatical phenomena cannot be identified in a corpus except by laborious manual intervention, and there is a temptation to focus on matters which can be dealt with more easily. It is a major problem that prosody, one of the most basic aspects of speech, is totally ignored. In a number of cases, e.g. in connection with word order, the analysis is based on a small selection of texts from the corpus. The reader is strongly recommended to consult the analysis notes referred to in the 'corpus findings' sections and printed at the end of the book.

A simple example will serve to illustrate problems in interpreting the quantitative findings. A comparison of the modal auxiliaries and the semi-modals shows that they are most common in conversation (p. 486). In the discussion of the findings it is said that "the greater frequency of both modals and semi-modals is understandable given that these forms mostly convey stance-type meanings" (p. 487), and a cross-reference is made to Chapter 12 (on the grammatical marking of stance). What one must keep in mind, however, is that clauses are shorter and more numerous in conversation than in the written registers, as evidenced by the high proportion of verbs (p. 65), and the more clauses there are, the more opportunities there are for using modal auxiliaries. Ideally, one would have liked to calculate the frequency of modals in relation to the number of clauses, not in relation to words per million. But this was unworkable in the absence of a syntactically parsed corpus (like the one produced by Ellegård; cf. Section **2.2**). The same point applies to a great many other quantitative observations; frequencies in terms of words per million are simply a matter of convenience, as they can be calculated more easily, but they may complicate the interpretation of the findings.

It is probably true, nevertheless, that modals are more common in conversation than in the written registers. It agrees with the more general frequency of stance markers (p. 979) as well as with Hermerén's findings (reported in Section **3**). These findings help in the interpretation of the decline of the modals referred to in Section **2.4**. Considering that the modals appear to be more characteristic of speech than of writing, it is unlikely that the recent change in their frequency is due to colloquialisation of written norms. If so, one would have expected an increase rather than a decline. But the modals are many-faceted and need to be differentiated by function and use, before definite conclusions can be drawn.[18]

[18] For further discussion of reasons for the decline of the modals, see Leech (2003, 2004), where there are observations on spoken as well as written corpora. Leech (2003) is also significant in including some discussion of semantic aspects of modal decline.

A further point which complicates the interpretation of the corpus findings in the grammar is that the register categories are very broad and that there is a great deal of variation within each. It can indeed be questioned whether the notion of register is applicable at all. It would have been preferable to make distinctions within the registers or, at least, have some measure of dispersion within the registers rather than just indications of mean frequencies. This would have made the task unmanageable. In spite of these problems, it should be possible, on the basis of the results we present, to get a general picture of how grammatical features vary across registers and to build up a profile defining the grammatical characteristics of each of the main registers. The results can be used as a yardstick in the study of grammatical features of texts and as a basis for further research, because much remains to be done in the area of corpus-based grammar studies.

6. Other developments

Below I will briefly refer to some other developments and their relevance with respect to a discussion of problems in interpreting textual distribution.

6.1 The ICE and ICLE families

The International Corpus of English (ICE) was launched by Sidney Greenbaum (1991), with the primary aim of collecting material for comparative studies of English worldwide, both used as a mother tongue and as a second language, and including both spoken and written texts, thus allowing for variation analysis both within and across corpora. Each ICE corpus was to consist of one million words of English produced after 1989 and collected in accordance with a common design (Nelson 1996). Seven ICE corpora are currently available.[19] It goes without saying that it is a great challenge to compile comparable English corpora for language communities as different as in East Africa, Great Britain, India, and Singapore, to mention but four of the participating regions. Leitner (1991) questions whether it is possible for English corpora to satisfy both a local requirement (true representation of texts from a particular area) and a global requirement (comparability across corpora from different areas). This is what we read about ICE Africa:[20]

> When compiling this corpus we followed the ICE stipulations as closely as possible to make a comparison with corpora of other varieties possible. At some points a few minor modifications of categories were necessary for two reasons:
>
> - the difficulty of acquiring data for some of the ICE categories
> - the linguistic situation in East Africa.
>
> For example, it was difficult to acquire a sufficient number of texts for the natural science category because there are simply not enough monographs of this sort written and published by East Africans. English radio programmes can be recorded in both countries [Kenya and Tanzania], but there are far fewer listeners in Tanzania because Swahili is the preferred language in every day conversations.

Another modification was an increase in the corpus size, to do justice to the two countries represented. Similar problems must have arisen in connection with other ICE projects. The

[19] These are the ICE corpora for East Africa, Great Britain (see **6.2**), Hong Kong, India, New Zealand, the Philippines, and Singapore. See: http://ice-corpora.net/ice/ (accessed in October 2012, HH)
[20] See: http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/eafrica/index.htm

material is no doubt extremely valuable, but the user should take great care in interpreting data from the corpora.

Similar caution is advised in the use of the subcorpora of *The International Corpus of Learner English* (ICLE), which is designed to collect data from English language learners from different parts of the world (Granger 1998). Although there is a common overall framework, learning situations inevitably vary in different countries making it difficult to compile comparable corpora. Again it is clear that the material is very valuable in opening up new avenues of research, but the user should pay close attention to the learner data included with the corpus and the accompanying description of the status of English in the different countries.

6.2 ICE-GB

The British component of ICE needs to be singled out especially, as it offers unique possibilities of intra-corpus comparison (see Nelson *et al*. 2002). Not only is it tagged and parsed (and post-edited!), it also comes with dedicated retrieval software which allows the user to carry out very complex searches combining linguistic, social, and situational variables. Nelson *et al*. (p. 257 ff.) explain how the corpus can be used to design experiments. Rather than first observing and later interpreting variation in a corpus, the researcher formulates specific hypotheses in advance and tests these with reference to the corpus, thereby reducing the element of interpretation. The hypothesis is either confirmed or rejected. In addition, ICE-GB can of course be used in the same way as any other corpus, with the added advantage that the delicacy of searches can be much greater. We can, for example, ask: In what situations are tag questions used most? Are they used more by women than men? It remains for the user to analyse the material and interpret the function of the tag questions, e.g. as used in the classroom and in legal cross-examination, two situations where tag questions are especially common.

A drawback of ICE-GB is that it is rather small, just one million words, so that it may not provide sufficient material, particularly if searches are very narrowly defined. A special problem follows from the detailed tagging and parsing. Users must familiarise themselves with the way terms and categories are defined (e.g. to recognise that *up* in *take up* is dealt with as an adverb, while 'particle' is used for other purposes). If not, they may draw unwarranted conclusions.

6.3 The British National Corpus

What ICE-GB lacks in size, it makes up in depth. The opposite is true of *The British National Corpus* (BNC), which is no doubt the most valuable of the publicly available English corpora (Aston and Burnard 1998). With its 100 million words, its large text samples and broad representation of text categories, including informal speech from different parts of the country, it is suited for many types of variation studies: speech vs. writing, types of written texts, types of spoken interaction (dialogue vs. monologue), types of speakers (age, sex, etc.), and so on. In a recent study of apology formulae Deutschmann (2006) shows how the social and situational categories of the BNC can be put to good use, reaching the conclusion that "the potential for future socio-pragmatic research is boundless" (p. 217).

As always there are a number of things the user should keep in mind. Using the social categories requires caution, as shown by Deutschmann (p. 208f.). While the dialogue material in the BNC lacks the academic bias of the London-Lund Corpus and is much larger and includes a wider range of speakers, it is less well transcribed and is often difficult to interpret.

Prosody can only be deduced indirectly through punctuation. There are inconsistencies in the transcription. For example, the sequence *must have* may appear in these forms: *must have, must've, must of*. There are many errors in word class tagging.[21] All in all, however, these problems do not detract from the great value of the corpus, as long as the user treats the material with care.

7. Future directions

In a recent paper Charles Meyer (2004) asks the provocative question 'Can you really study linguistic variation in linguistic corpora?' He puts this to the test with reference to a case study based on some of the ICE subcorpora, and he concludes:

> The study demonstrates that although certain kinds of language variation can be validly studied in a corpus such as ICE, other kinds of variation require different kinds of corpora. (quoted from the abstract, p. 339)

With this I whole-heartedly agree. It is the task of corpus users to evaluate whether the corpus is suitable for their research questions. All corpora are not suitable for all types of research questions. It is nevertheless striking how much research has come out of early corpora such as Brown and LOB, including investigations which the compilers had probably never even dreamed of.

Much earlier Nelleke Oostdijk (1988: 12) observed that "linguists have not been very well equipped to carry out large scale formal empirical analyses which would enable them to systematically vary extra-linguistic factors and examine the accompanying linguistic variation." Part of the problem with many corpus-based variation studies is that researchers have unquestioningly accepted genre categories that have been built into publicly available corpora. Oostdijk asks for a new type of corpus compiled for the specific purpose of variety study and allowing for the systematic varying of extra-linguistic variables.

With time corpora have become more numerous and more varied, opening up new opportunities for research. In addition to multipurpose corpora such as the BNC, we have specialised corpora of textbooks, academic writing, learner language, etc.[22] Problems of interpretation are reduced when the corpus user can choose the best corpus for a particular research question. Corpus annotation and new analysis software aid in the processes of investigation and interpretation. What must not be forgotten is that documentation is vital for the use of the corpora and for the interpretation of the findings: information on the texts, on text categories, on speakers, the kinds of annotation, etc. The BNC sets a good example.

Research questions have frequently been tailored according to the available ready-made corpora. With the ever-increasing material in electronic form, it is becoming easier to tailor the corpus to the research question. Many have started to talk of the Web as a corpus (see e.g. Kilgarriff 2001b and Renouf 2003). I would rather view it as a vast text archive from which different types of corpora can be drawn depending upon the research question. In a recent paper Christian Mair (2006) argues for the use of material from the Web:

[21] Note, incidentally, that the word class tagging differs from that of ICE-GB, not just in the level of delicacy but in the way certain word classes are defined.

[22] Among important types of corpora which I have not discussed, we find multilingual corpora such as the English-Norwegian Parallel Corpus (see: http://www.hf.uio.no/ilos/english/services/omc/enpc/) and historical corpora (see e.g.: http://www.eng.helsinki.fi/varieng/). For a discussion of some problems in using historical corpora, see Rissanen (1989).

> Today […] the supply of digital text – online and offline – is practically unlimited for English and a small number of other languages, so that restricting the scope of one's work to data available in a small number of corpora only would be counter-productive in the analysis of many linguistic problems. Corpus linguists of the future will therefore be a much more heterogeneous community – no longer focussed on a specific corpus but working in a vast and expanding corpus-linguistic environment in which one of the chief skills required will be to identify the resources which are relevant to the problem studied from a vast range of possibilities. (p. 370)

In using the Web it may be difficult to evaluate the material and indeed also to interpret the findings. Christian Mair discusses some cautionary procedures which need to be followed. In my view, the Web is *one* possible source, to be used where it is applicable.[23] The guiding principle is the nature of the research question.

Like many others, Christian Mair talks about 'corpus linguists', and 'corpus linguistics' has indeed become a household word.[24] The terms have been significant in underlining that using corpora is an important new undertaking, but they may have contributed to creating a gulf in relation to other types of linguistic inquiry, such as sociolinguistics. In preparing this paper, I turned to the *Handbook of language variation and change* (Chambers *et al*. 2002). Although this is a volume of about 800 pages, I found just one contribution on corpora (Bauer 2002). Yet 'corpus linguistics' and sociolinguistics are not incompatible. A recent example is a study of 'Syntactic variation and beyond: Gender and social class variation in the use of discourse-new markers' (Cheshire 2005), using a speech corpus assembled for a sociolinguistic project where social and situational factors were strictly controlled. In dealing with information structure, she enters an area which has occupied corpus researchers, and she refers in passing to work on the London-Lund Corpus (p. 481).

Cheshire reaches the conclusion that it is necessary to go beyond conventional sociolinguistic frameworks of analysis and she "confirms the view articulated by Pintzuk (2003: 525) that a coherent theory relating grammar and usage can and should be formulated" (p. 502). Although much new insight on variation has been gained through corpus studies, leading to a better understanding of texts and of the use of language, there is a need to go beyond customary 'corpus linguistic' approaches to variation studies. Now is an appropriate time for linguists of different persuasions to join forces to advance our understanding of language in use as well as for the development of linguistic theory and new models of language variation.

References

Aijmer, Karin and Bengt Altenberg (eds). 1991. *English corpus linguistics. Studies in honour of Jan Svartvik*. London: Longman.

Altenberg, Bengt. 1984. Causal linking in spoken and written English. *Studia Linguistica* 38: 20-69.

Altenberg, Bengt. 1986. Contrastive linking in spoken and written English. In: Tottie and Bäcklund (1986), 13-40.

Aston, Guy and Lou Burnard. 1998. *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press

[23] Cf. Bergh (2005: 45): "the Web is best enjoyed in carefully cut slices, preferably based on the raw capacity of Google and spiced according to taste with the fine-tuned facilities of WebCorp".

[24] One of those who have been sceptical of the use of these terms is Wallace Chafe (1992).

Bäcklund, Ingegerd. 1986. Beat until stiff. Conjunction-headed abbreviated clauses in spoken and written English. In: Tottie and Bäcklund (1986), 41-55.
Ball, C. N. 1994. Automated text analysis: Cautionary tales. *Literary and Linguistic Computing* 9 (4): 295-302.
Bauer, Laurie. 2002. Inferring variation and change from public corpora. In: Chambers *et al*. (2002), 97-114.
Bergh, Gunnar. 2005. Min(d)ing English language data on the Web: What can Google tell us? *ICAME Journal* 29: 25-46.
Biber, Douglas. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language* 62 (2): 384-414.
Biber, Douglas 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.
Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.
Biber, Douglas and Edward Finegan. 1991. On the exploitation of computerized corpora in variation studies. In: Karin Aijmer and Bengt Altenberg (1991), 204- 220.
Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
Biber, Douglas, Stig Johansson, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.
Carroll, John B., Peter Davies, and Barry Richman. 1971. *The American Heritage word frequency book*. New York: American Heritage Publishing Co.
Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In: Deborah Tannen (ed.), *Spoken and written language: Exploring orality and literacy*, 35-53. Norwood, N.J.: Ablex.
Chafe, Wallace. 1992. The importance of corpus linguistics to understanding the nature of language. In: Jan Svartvik (ed.), *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, 4-8 August 1991*, 79-103. Berlin: Mouton de Gruyter.
Chambers, J. K., Peter Trudgill, and Natalie Schilling-Estes (eds). 2002. *The handbook of language variation and change*. Oxford: Blackwell.
Cheshire, Jenny. 2005. Syntactic variation and beyond: Gender and social class variation in the use of discourse-new markers. *Journal of Sociolinguistics* 9 (4): 479-508.
Collins, Peter. 1991. *Cleft and pseudo-cleft constructions in English*. London and New York: Routledge.
Collins, Peter and Pam Peters. 2004. Australian English: Morphology and syntax. In: Kortmann (2004), 593-610.
Deutschmann, Mats. 2006. Social variation in the use of apology formulae in the BNC. In: Renouf and Kehoe (2006), 205-221.
Ellegård, Alvar. 1978. *The syntactic structure of English texts. A computer-based study of four kinds of text in the Brown University Corpus*. Gothenburg Studies in English 43. Göteborg: Acta Universitatis Gothoburgensis.
Esser, Jürgen. 1993. *English linguistic stylistics*. Tübingen: Niemeyer.
Francis, W. Nelson and Henry Kučera (with the assistance of Andrew W. Mackie). 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.
Granger, Sylviane (ed.). 1998. *Learner English on computer*. London and New York: Longman.
Greenbaum, Sidney. 1991. The development of the International Corpus of English. In: Aijmer and Altenberg (1991), 83-91.
Greenbaum, Sidney and Jan Svartvik. 1990. The London-Lund Corpus of Spoken English. In:

Svartvik (1990), 11-59.
Hermerén, Lars. 1986. Modalities in spoken and written English: An inventory of forms. In: Tottie and Bäcklund (1986), 57-91.
Hofland, Knut and Stig Johansson. 1982. *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities and London: Longman.
Holmes, Janet and Robert Sigley. 2002. What's a girl doing in a place like this? Occupational labels, sexist usages and corpus research. In: Pam Peters, Peter Collins, and Adam Smith (eds), *New frontiers of corpus research. Papers from the Twenty First International Conference on English Language Research on Computerized Corpora*, 247-263. Amsterdam and New York: Rodopi.
Hundt, Marianne. 1997. Has BrE been catching up with AmE over the past thirty years? In: Ljung (1997), 135-151.
Hundt, Marianne, Jennifer Hay, and Elizabeth Gordon. 2004. New Zealand English. In: Kortmann *et al*. (2004), 560-592.
Johansson, Stig (in collaboration with Geoffrey Leech and Helen Goodluck). 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers*. Department of English, University of Oslo.
Johansson, Stig and Knut Hofland. 1989. *Frequency analysis of English vocabulary and grammar*. Vol. 1-2. Oxford: Clarendon Press.
Johansson, Stig and Anna-Brita Stenström (eds). 1991. *English computer corpora: Selected papers and research guide*. Berlin: Mouton de Gruyter.
Jones, Susan and John M. Sinclair. 1974. English lexical collocations. *Cahiers de Lexicologie* 24: 15-61.
Kilgarriff, Adam. 2001a. Comparing corpora. *International Journal of Corpus Linguistics* 6 (1): 1-37.
Kilgarriff, Adam 2001b. Web as corpus. In: Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie, and Shereen Khoja (eds), *Proceedings of the Corpus Linguistics 2001 conference*, 342-344. Lancaster: UCREL.
Kortmann, Bernd, Kate Burridge, Rajend Mesthrie, Edgar W. Schneider, and Clive Upton. 2004. *A handbook of varieties of English*. Vol. 2: *Morphology and syntax*. Berlin and New York: Mouton de Gruyter.
Krogvig, Inger and Stig Johansson. 1984. *Shall* and *will* in British and American English: A frequency study. *Studia Linguistica* 38 (1): 70-87.
Kučera, Henry and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Providence, R.I.: Brown University Press.
Labov, William. 1966. *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Leech, Geoffrey. 2003. Modality on the move: The English modal auxiliaries 1961-1992. In: Roberta Facchinetti, Manfred Krug, Frank Palmer (eds), *Modality in contemporary English*, 223-240. Berlin and New York: Mouton de Gruyter.
Leech, Geoffrey. 2004. Recent grammatical change in English: Data, description, theory. In: Karin Aijmer and Bengt Altenberg (eds), *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23), Göteborg 22-26 May 2002*, 61-81. Amsterdam and New York: Rodopi.
Leech, Geoffrey and Roger Fallon. 1992. Computer corpora – what do they tell us about culture? *ICAME Journal* 16: 29-50.

Leech, Geoffrey and Nicholas Smith. 2005. Extending the possibilities of corpus-based research on English in the twentieth century: A prequel to LOB and FLOB. *ICAME Journal* 29: 83-98.

Leitner, Gerhard. 1991. The Kolhapur Corpus of Indian English. Intra-varietal description and/or inter-varietal comparison. In: Johansson and Stenström (1991), 215-232.

Ljung, Magnus (ed.). 1997. *Corpus-based studies in English. Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17)*. Amsterdam and Atlanta: Rodopi.

Lyons, John. 1968. *Introduction to theoretical linguistics*. London: Cambridge University Press.

Mair, Christian. 1997. Parallel corpora: A real-time approach to the study of language change in progress. In: Ljung (1997), 195-209.

Mair, Christian. 1998. Corpora and the study of the major varieties of English: Issues and results. In: Hans Lindquist, Staffan Klintborg, Magnus Levin, and Maria Estling (eds), *The major varieties of English. Papers from MAVEN 97, Växjö 20-22 November 1997*, 139-157. Acta Wexionensia Humaniora. No. 1.Växjö: Växjö University.

Mair, Christian. 2006. Tracking ongoing change and recent diversification in present-day standard English: The complementary role of small and large corpora. In: Renouf and Kehoe (2006), 355-376.

Mair, Christian and Marianne Hundt. 2002. Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics* 7 (2): 245-264.

Meyer, Charles F. 2004. Can you really study linguistic variation in linguistic corpora? *American Speech* 79 (4): 339-355.

Nakamura, Junsaku. 1993. Quantitative comparison of modals in the Brown and LOB corpora. *ICAME Journal* 17: 29-48.

Nelson, Gerald. 1996. The design of the corpus. In: Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 27-35. Oxford: Clarendon.

Nelson, Gerald, Sean Wallis, and Bas Aarts. 2002. Exploring natural language. *Working with the British component of the International Corpus of English*. Amsterdam and Philadelphia: Benjamins.

Oostdijk, Nelleke. 1988. A corpus linguistic approach to linguistic variation. *Literary and Linguistic Computing* 3 (1): 12-25.

Pintzuk, Susan. 2003. Variationist approaches to syntactic change. In: Brian D. Joseph and Richard D. Janda (eds), *Handbook of historical linguistics*, 506-528. Oxford: Blackwell.

Quirk, Randolph. 1960. Towards a description of English usage. *Transactions of the Philological Society, 1960*, 40-61. Oxford: Blackwell.

Renouf, Antoinette. 2003. WebCorp: Providing a renewable data source for corpus linguists. *Language and Computers* 48: 39-58.

Renouf, Antoinette and Andrew Kehoe (eds). 2006. *The changing face of corpus linguistics*. Amsterdam and New York: Rodopi.

Reppen, Randi, Susan M. Fitzmaurice, and Douglas Biber. 2002. *Using corpora to explore linguistic variation*. Amsterdam and Philadelphia: Benjamins.

Rissanen, Matti. 1989. Three problems in connection with the use of diachronic corpora. *ICAME Journal* 13: 16-19.

Sand, Andrea and Rainer Siemund. 1992. LOB – 30 years on … , *ICAME Journal* 16: 119-122.

Shastri, S. V. 1988. The Kolhapur Corpus of Indian English and work done on its basis so far. *ICAME Journal* 12: 15-26.

Sigley, Robert. 1997. The influence of formality and channel on relative pronoun choice in New Zealand English. *English Language and Linguistics* 1 (2): 207-232.
Svartvik, Jan (ed.). 1990. *The London-Lund Corpus of Spoken English: Description and research*. Lund Studies in English 82. Lund: Lund University Press.
Svartvik, Jan and Randolph Quirk (eds).1980. *A corpus of English conversation*. Lund Studies in English 56. Lund: CWK Gleerup.
Tottie, Gunnel. 1986. The importance of being adverbial. Adverbials of focusing and contingency in spoken and written English. In: Tottie and Bäcklund (1986), 93-118.
Tottie, Gunnel. 1991. *Negation in English speech and writing*. San Diego etc.: Academic Press.
Tottie, Gunnel and Ingegerd Bäcklund (eds). 1986. *English in speech and writing: A symposium*. Studia Anglistica Upsaliensia 60. Uppsala: Almqvist & Wiksell.
Wilson, Andrew. 2005. Modal verbs in written Indian English: A quantitative and comparative analysis of the Kolhapur Corpus using correspondence analysis. *ICAME Journal* 29: 151-169.